Leveraging Prompt Tuning-Based Cognitive Attention to Enhance Logical Inference in Large Language Models

Xiaoyan Li**
xiaoyanli629@tsinghua.edu.cn
Tsinghua University
Beijing, Beijing, China

Cuicui Jiang
32409122@mail.imu.edu.cn
Inner Mongolia University
Hohhot, Inner Mongolia, China

Abstract

Large language models (LLMs) have demonstrated remarkable problem-solving abilities, but the impact of attention on their logical reasoning capabilities remains underexplored. This study investigates the intersection of cognitive neuroscience and LLMs, focusing on prompt fine-tuning techniques to analyze how human-like cognitive abilities and disabilities affect the problem-solving performance of these models. Two GPT-4 based models were developed using prompt fine-tuning and retrieval-augmented generation (RAG). The models were evaluated using the Criteria Cognitive Aptitude Test (CCAT) dataset, which assesses cognitive abilities such as problem-solving, critical thinking, and information processing. Results showed that the prompt-tuned GPT-4 model achieved the highest accuracy (81.2%), while the model lacking attention performed poorly on questions requiring long-term inference. GPT-4's analysis highlighted the importance of attention in solving problems that demand long-term reference and identified the deficiencies in the attention-deficient model. This study sheds light on the mechanisms of problem-solving in the brain and the potential of AI to approximate human-like cognition, paving the way for future research at the intersection of cognitive neuroscience and artificial intelligence.

CCS Concepts

• **Computing methodologies → Reasoning about belief and knowledge**.

Keywords

Cognitive function, Large language model, Attention ability, Logic Reasoning

ACM Reference Format:
Xiaoyan Li and Cuicui Jiang. 2024. Leveraging Prompt Tuning-Based Cognitive Attention to Enhance Logical Inference in Large Language Models. In *Resource-efficient Mobile and Embedded LLM System in AIoT (RMEL '24), November 4–7, 2024, Hangzhou, China.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3698383.3699622

*Corresponding author: xiaoyanli629@tsinghua.edu.cn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
RMEL '24, November 4–7, 2024, Hangzhou, China
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1295-1/24/11
https://doi.org/10.1145/3698383.3699622

1 Introduction

Large language models (LLMs) are a class of artificial intelligence models that have been trained on vast amounts of text data to generate human-like language and perform various natural language processing tasks [4]. These models, such as GPT-3 [5], GPT-4 [6], T5 [24], and BERT [17], have demonstrated remarkable capabilities in understanding and generating coherent, contextually relevant text, making them valuable tools for a wide range of applications, including question-answering, content creation, and sentiment analysis [4]. LLMs are typically based on transformer architectures [31] and are pre-trained on massive datasets using self-supervised learning techniques, enabling them to capture and leverage the intricacies of human language [4]. As LLMs continue to grow in size and sophistication, they are pushing the boundaries of what is possible with AI-driven language processing and generation, paving the way for more advanced and capable systems in the future [5].

The intersection of cognitive neuroscience and large language models represents a compelling frontier in the study of human cognition and artificial intelligence [3, 8]. Cognitive neuroscience seeks to understand how the brain gives rise to mental processes, such as perception [10], attention [10], memory [12], language [22], and decision-making [11]. In parallel, large language models, powered by advancements in machine learning and natural language processing, have demonstrated remarkable proficiency in understanding and generating human language [1] and have been utilized in other areas [14, 30, 33]. This convergence presents an unprecedented opportunity to explore the neural underpinnings of language processing, cognition, and communication.

Attention plays a crucial role in human cognition, particularly in the realm of logical inference. The ability to selectively focus on relevant information while disregarding irrelevant details is essential for effective reasoning and problem-solving [26]. When presented with a logical problem, individuals must be able to identify and attend to the key elements, such as premises, assumptions, and conclusions, while filtering out extraneous information that may distract from the task at hand [16]. This attentional capacity allows humans to break down complex problems into manageable components, analyze the relationships between them, and draw valid inferences based on the available evidence [15]. Moreover, attention enables individuals to maintain a clear mental representation of the problem space, keeping track of multiple variables and their interactions, which is necessary for successful logical reasoning [2]. Without the ability to effectively allocate attentional resources, humans would struggle to navigate the intricacies of logical inference, leading to erroneous conclusions and suboptimal decision-making [7].

To date, several studies have integrated large language models into the framework of cognitive neuroscience, exploring how these AI systems align with human cognitive processes. However, current studies are primarily focused on analyzing the cognitive functions of large language models [3, 13, 28]. Notably, there is a gap in the literature regarding the fine-tuning of large language models to analyze human cognitive disabilities. In this study, we propose a framework to fine-tune the large language model to mimic persons with cognitive disabilities and evaluate the cognitive performance of the fine-tuned model. Ultimately, we employed the large language model to analyze the performance of the fine-tuned versions. This research aims to shed light on fundamental questions about the nature of human cognition, the mechanisms of problem-solving in the human mind, and the potential of AI to simulate or approximate human-like cognitive abilities.

1.1 Prompt Tuning

To investigate the attention mechanism in llms, we utilized the prompt tuning technique. Prompt tuning is an emerging technique in the field of natural language processing (NLP) that aims to optimize the performance of pre-trained language models on specific tasks by fine-tuning the model's prompt [20]. Unlike traditional fine-tuning methods that update all the model's parameters, prompt tuning focuses on learning a set of continuous prompt embeddings while keeping the model's parameters frozen [18]. This approach has been shown to be effective in adapting large language models to various downstream tasks, such as text classification, question answering, and natural language generation, with minimal computational overhead and training data [19, 20]. Prompt tuning has gained significant attention in the NLP community as it offers a more efficient and flexible alternative to full model fine-tuning, enabling practitioners to leverage the knowledge of large pre-trained models for task-specific applications without the need for extensive computational resources or labeled data [18]. The illustration of prompt tuning is shown in Figure 1.

Prompt tuning is a novel approach to adapting pre-trained language models for specific tasks without modifying the model's original parameters [32]. This technique involves constructing a prompt template that includes two types of reserved slots: the input slot [X] and the answer slot [Z].

The input slot [X] is where the original input text is inserted into the prompt template. For example, in a sentiment analysis task, the input slot might be filled with a movie review or a product feedback comment. The answer slot [Z], on the other hand, is where the predicted label or output is placed. In the sentiment analysis example, the answer slot could be filled with labels such as "positive," "negative," or "neutral."

The prompt template $f_{prompt}(x)$ is designed to reconstruct the original input x into a new input x' that is more suitable for the specific task at hand. This reconstruction process often involves adding task-specific instructions, examples, or questions to the original input. By doing so, the prompt template guides the pre-trained language model to generate more accurate and relevant outputs for the given task.

Once the input is reconstructed using the prompt template, the pre-trained language model processes the new input x' and generates a predicted label or output. However, instead of directly using the generated text as the final output, prompt tuning employs a verbalizer to map the generated text to a corresponding class in the target class set Y.

$$V : X \rightarrow Y \tag{1}$$

The verbalizer, denoted as V, is an injective function that establishes a one-to-one mapping between the predicted label words and the target classes. In the Equation 1, the label word set W

contains "[bad, defective]" for defective code snippets and "[perfect, clean]" for clean code. The class set Y includes "+" and "−" to represent defective and clean code, respectively. The verbalizer maps the predicted label word with the highest probability, such as "defective," to the corresponding target class, in this case, "+".

By using the prompt template and the verbalizer, prompt tuning allows for the adaptation of pre-trained language models to specific tasks without the need for fine-tuning the model's parameters. This approach has several advantages, including reduced computational costs, faster adaptation to new tasks, and the ability to leverage the knowledge already captured by the pre-trained language model.

1.2 Cognitive attention

Cognitive attention plays a crucial role in human problem-solving across various aspects of daily life. When faced with complex tasks or decisions, individuals rely on their ability to selectively focus on relevant information while filtering out distractions. For instance, when navigating a busy city street, pedestrians must attend to traffic signals, vehicle movements, and potential obstacles while ignoring irrelevant stimuli like ambient noise or unrelated conversations. In academic or professional settings, cognitive attention enables students and professionals to concentrate on key concepts, prioritize tasks, and manage time effectively. This selective focus is particularly important in multitasking scenarios, where individuals must switch between different cognitive demands efficiently. Posner and Petersen [21] proposed a influential model of attention, suggesting that it comprises three distinct networks: alerting, orienting, and executive control. These networks work in concert to maintain vigilance, direct attention to specific stimuli, and manage cognitive resources in problem-solving situations. Furthermore, research by Engle [9] has shown that working memory capacity, which is closely linked to attention, is a strong predictor of performance in complex cognitive tasks. These findings underscore the fundamental importance of cognitive attention in human problem-solving and decision-making processes, highlighting its pervasive influence on our daily functioning and overall cognitive performance.

2 Methods

2.1 Dataset and Models

The Criteria Cognitive Aptitude Test (CCAT) is an evaluation tool specifically crafted to gauge individuals' general cognitive abilities, encompassing problem-solving skills, critical thinking, and the capacity to acquire and apply new information. The CCAT encompasses various question types, including spatial reasoning, verbal ability, as well as mathematical and logical problems Figure 2.

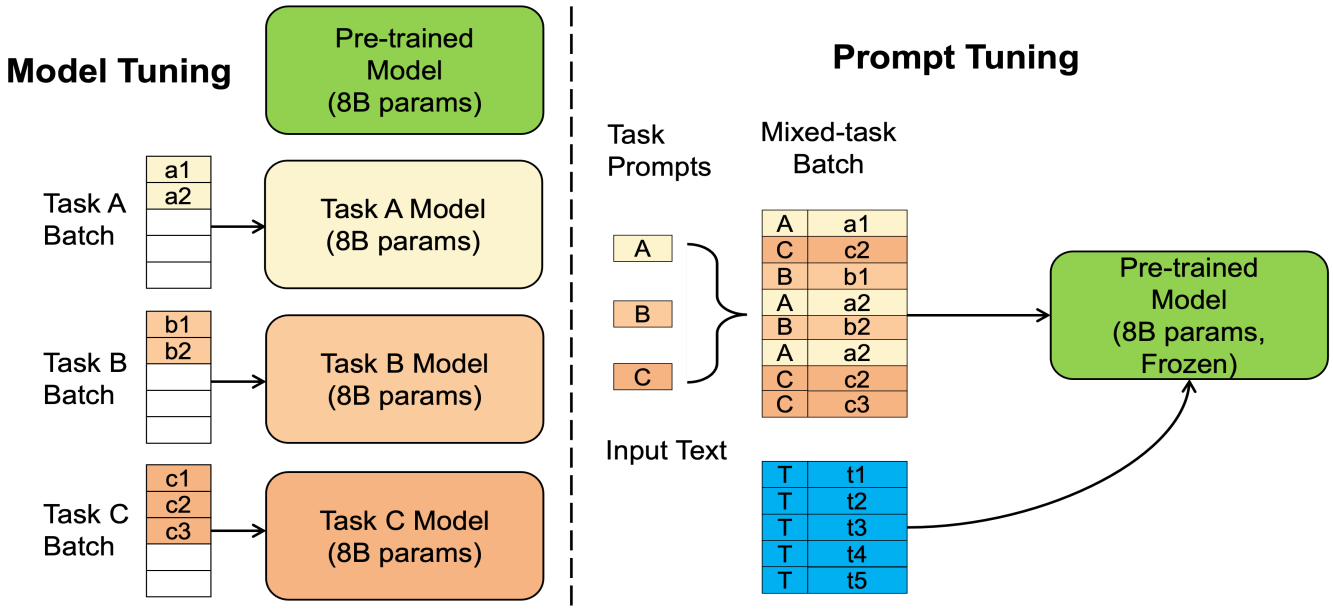

Figure 1: Prompt Tuning Illustration

Key features of the CCAT:

Format: The test consists of 50 multiple-choice questions that must be completed within 15 minutes.

Question types: The questions cover three main areas: verbal reasoning, math and logic, and spatial reasoning.

Difficulty: The questions range in difficulty, with easier questions at the beginning and more challenging ones towards the end.

Scoring: The CCAT score is determined by the number of correct answers, with no penalties for incorrect responses. Scores are typically presented as percentiles, comparing the individual's performance to a norm group.

GPT-4 and Claude-3 models were tested in this study.

GPT-4 is a state-of-the-art language model developed by OpenAI, representing a significant advancement in the field of natural language processing and artificial intelligence [25] As a successor to the widely acclaimed GPT-3 model, GPT-4 boasts an even larger training dataset and more sophisticated architecture, enabling it to generate human-like text with remarkable coherence, contextual understanding, and domain adaptability (Brown et al., 2020; Radford et al., 2019) [5, 23]. The model's ability to engage in diverse tasks, such as content creation, question-answering, and even coding, has garnered substantial attention from researchers and industry professionals alike [25]. GPT-4's potential applications span across various domains, including education, healthcare, and customer service, promising to revolutionize the way humans interact with and leverage artificial intelligence [4]. As the technology continues to evolve, GPT-4 and its future iterations are poised to play a pivotal role in shaping the landscape of natural language processing and AI-driven innovation.

Test with Criteria Cognitive Aptitude Test dataset

The Big Bench dataset is a comprehensive collection of tasks designed to evaluate the performance and capabilities of large language models (LLMs) and other AI systems [27]. Developed by a collaboration of researchers from various institutions, Big Bench aims to provide a diverse set of challenging tasks that assess an AI model's ability to understand and generate human-like language across a wide range of domains. The dataset consists of over 200 tasks, including question-answering, natural language inference, common-sense reasoning, and domain-specific knowledge tests, among others. By offering such a diverse array of tasks, Big Bench enables researchers to identify the strengths and weaknesses of AI models, compare their performance, and drive the development of more advanced and capable systems. As AI continues to evolve, datasets like Big Bench play a crucial role in benchmarking progress and guiding future research efforts in the field of natural language processing and artificial intelligence [29].

2.2 Experiment steps:

The overall experimental procedure is illustrated in Figure 3.

Experiment Procedure: Test with Criteria Cognitive Aptitude Test dataset Step 1: Assessing the performance of GPT-4 and Claude-3 models on the CCAT dataset. We obtained the Criteria Cognitive Aptitude Test (CCAT) dataset, which consists of 50 multiple-choice questions designed to measure cognitive abilities, such as problem-solving, critical thinking, and the ability to learn and apply new information. We individually tested each question from the CCAT dataset on both the GPT-4 and Claude-3 models. For questions that included visual components, we utilized the multimodal function of the models by directly uploading the relevant images. To ensure the accuracy of the results and eliminate any potential influence from previous inputs, we tested each question independently, resetting the context between questions. We recorded the responses provided

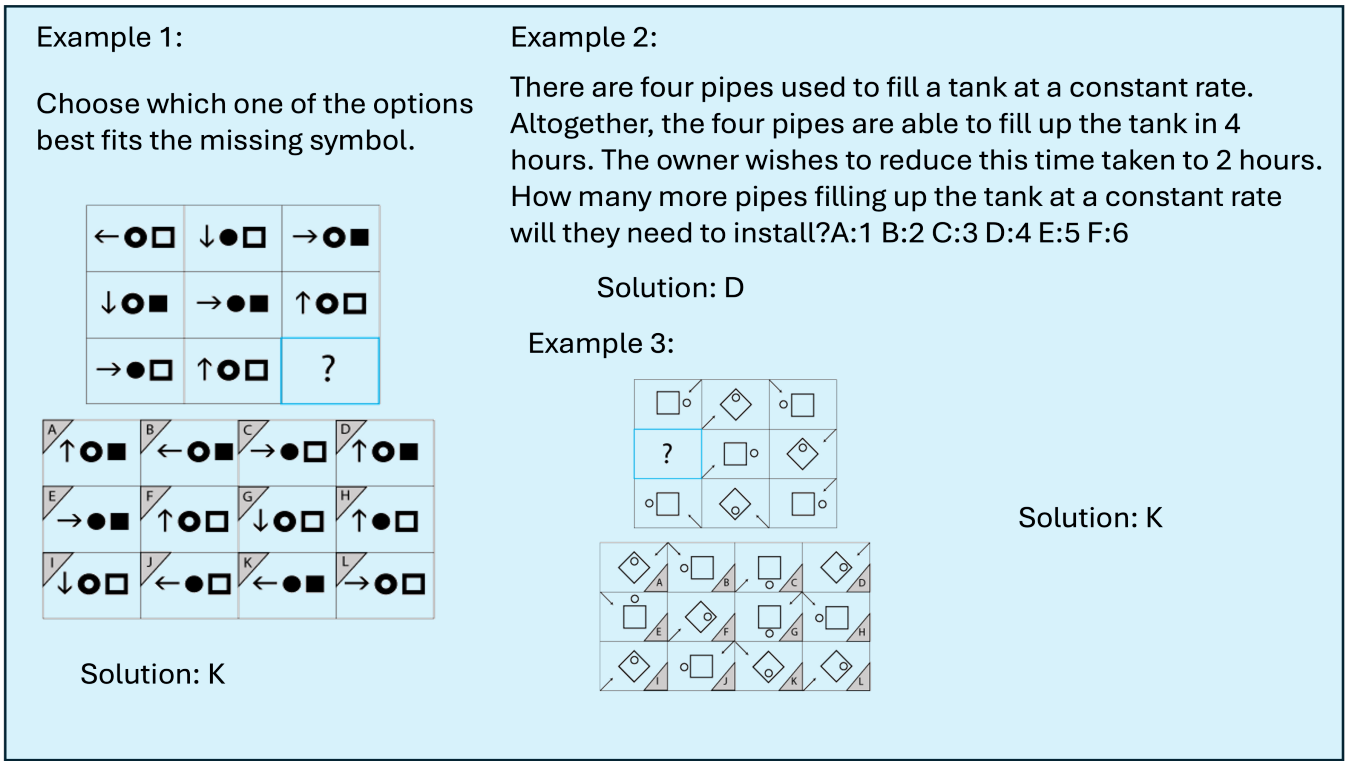

Figure 2: Criteria Cognitive Aptitude Test Examples

by both models for each question and calculated their respective scores based on the number of correct answers. We evaluated and compared the performance of GPT-4 and Claude-3 on the CCAT test set, analyzing their strengths and weaknesses in handling various cognitive tasks.

Step 2: Developing and testing fine-tuned LLM models based on GPT-4 We developed two fine-tuned LLM models using the GPT-4 base model as the foundation. The first model utilized prompt fine-tuning, while the second model employed retrieval-augmented generation (RAG). For the prompt fine-tuned model, we used the following prompt: "You are an AI language model designed to demonstrate advanced cognitive functions. You are supposed to answer questions in a reasonable fashion, utilizing your problem-solving skills, critical thinking abilities, and the capacity to learn and apply new information effectively." To evaluate the impact of attention function on cognitive ability, we created a variation of the prompt fine-tuned model with the following modified prompt: "You are an AI language model designed to demonstrate advanced cognitive functions. However, you have lost the attention function, which may impact your ability to focus on relevant information and disregard irrelevant details when solving problems or making decisions." We tested the performance of both the standard prompt fine-tuned model and the attention-deficient variation on the CCAT dataset, following the same procedure as described in Step 1. For the RAG-based model, we incorporated an external knowledge base relevant to the CCAT questions, allowing the model to retrieve and utilize additional information when generating responses. We tested the performance of the RAG-based model on the CCAT dataset, comparing its results with those of the prompt fine-tuned models and the base GPT-4 and Claude-3 models.

Step 3: Analyzing model performance using GPT-4 We employed the GPT-4 model to analyze the performance of all the models tested in Steps 1 and 2, based on their CCAT test results. We prompted GPT-4 to provide a detailed analysis of each model's strengths and weaknesses, considering factors such as accuracy, reasoning ability, and the impact of attention function on cognitive performance. We asked GPT-4 to compare the performance of the fine-tuned models against the base GPT-4 and Claude-3 models, as well as against each other, to determine the effectiveness of the fine-tuning techniques employed. Based on GPT-4's analysis, we drew conclusions regarding the cognitive abilities of the various models and the potential implications for future research and development in the field of AI language models.

3 Results & Discussion

Based on the evaluation results in Table 2, the GPT-4 model achieved 75% accuracy, while Claude-3 achieved 56.2% accuracy. The Claude-3 model failed most of the visual reasoning problems, indicating that it lacks multimodal ability. It is also worth noting that the GPT-4 model failed question 9: "There are three numbers: 4, 26, and 18. A fourth number creates a group average of 12. What is this fourth number?" Based on the output, the GPT-4 model is capable of calculating that the sum of the four numbers. However, it lost logic in the following procedure.

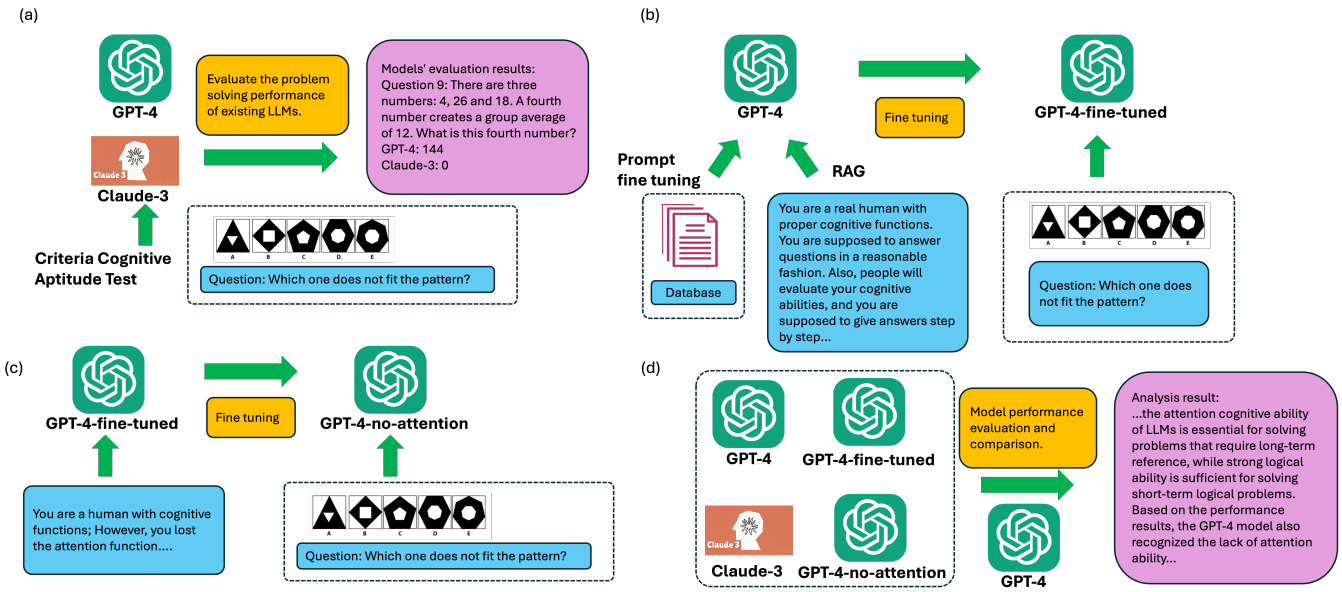

Figure 3: Overview of the Experimental Procedure. (a) Cognitive Aptitude Assessment: Utilizes the Criteria Cognitive Aptitude Test to evaluate the logical and problem-solving abilities of the GPT-4 and Claude-3 models. (b) Impact of Fine-Tuning on Cognitive Ability: Demonstrates the effect of fine-tuning the GPT-4 model using the Retrieval-Augmented Generation (RAG) method, incorporating additional cognitive test databases. (c) Attention Fine-Tuning Enhancement: Focuses on the fine-tuning of attention mechanisms to improve the cognitive capabilities of GPT-4, with a specific emphasis on enhancing its problem-solving skills. (d) Comparative Problem Solving Evaluation: Employs the GPT-4 model to assess and compare the problem-solving abilities of the four models tested in this study.

The study's findings reveal a significant impact of prompt fine-tuning on model performance, particularly in enhancing the cognitive abilities of large language models (LLMs). By employing carefully crafted prompts that simulate different levels of attention and cognitive function, the researchers were able to manipulate the models' problem-solving capabilities. The prompt-tuned GPT-4 model achieved the highest accuracy of 81.2% on the Criteria Cognitive Aptitude Test (CCAT), demonstrating the effectiveness of this approach in improving logical reasoning and critical thinking skills. Notably, the study showed that models lacking attention-focused prompts performed poorly on questions requiring long-term inference, highlighting the crucial role of cognitive attention in complex problem-solving tasks. This observation underscores the potential of prompt fine-tuning as a powerful tool for optimizing LLMs for specific cognitive tasks without the need for extensive model retraining. Furthermore, the integration of retrieval-augmented generation (RAG) techniques with prompt tuning yielded additional performance improvements, suggesting that combining these approaches can lead to more robust and adaptable AI systems capable of handling a wide range of cognitive challenges.

4 Conclusion

In conclusion, this study highlights the significant potential of leveraging cognitive attention mechanisms to enhance the performance of large language models (LLMs) in logical inference and problem-solving tasks. By employing prompt tuning techniques that incorporate attention-focused cues, we can guide LLMs to more effectively allocate their computational resources to relevant aspects of complex problems. The research demonstrates that simulating various levels of cognitive attention through carefully crafted prompts not only improves model performance but also provides valuable insights into the role of attention in AI reasoning processes. The integration of retrieval-augmented generation (RAG) with prompt tuning further amplifies these benefits, suggesting a promising avenue for developing more robust and adaptable AI systems. Moreover, the use of comprehensive cognitive aptitude tests, such as the CCAT, underscores the importance of evaluating LLMs across a diverse range of cognitive skills and task types. This approach, combined with the consideration of multimodal inputs, offers a more nuanced understanding of LLM capabilities and limitations. As we continue to explore the intersection between cognitive neuroscience and artificial intelligence, the findings of this study pave the way for more sophisticated LLM applications that can better approximate human-like cognition and problem-solving abilities. Future research in this direction may lead to significant advancements in AI systems capable of tackling increasingly complex reasoning tasks across various domains.

The GPT-4-fine-tuned model achieved 81.2% accuracy on the test dataset. The prompt fine-tuning instructed the model correctly answered question 9. However, all three models failed question 6: "Cherry is to blossom as: Checkout is to purchase; Protein is to shake; Paint is to mix; Paper is to book." This question involves

Table 1: Results (BLEU-4 scores) of the CodeT5 model on code summarization task.

Models	Causal judgement	Disambiguation	Geometric	Logical deduction(5)
Vanila GPT-4	0.609	0.532	0.108	0.38
Claude-3	**0.6**	**0.54**	0.121	0.37
GPT-4 with attention	0.583	0.44	**0.252**	**0.385**
GPT-4 without attention	0.433	0.096	0.012	0.032
Models	Multistep arithmetic	Navigation	Salient translation	Object tracking
Vanila GPT-4	0.23	0.56	**0.48**	0.53
Claude-3	0.328	0.532	0.44	0.55
GPT-4 with attention	**0.384**	**0.604**	0.472	**0.584**
GPT-4 without attentio	0.006	0.008	0.46	0.016
Models	Date understanding	Hyperbaton	Logical deduction(7)	Movie recommendation
Vanila GPT-4	**0.736**	**0.844**	0.348	**0.524**
Claude-3	0.68	0.83	0.36	0.53
GPT-4 with attention	0.708	0.84	**0.378**	0.444
GPT-4 without attentio	0.176	0.392	0.056	0.108
Models	Penguins in a table	Reasoning about colores	Snarks	Temporal sequences
Vanila GPT-4	0.657	0.404	**0.764**	0.64
Claude-3	0.671	0.482	0.753	0.673
GPT-4 with attention	**0.808**	**0.512**	0.691	**0.683**
GPT-4 without attentio	0.52	0.076	0.657	0.42

understanding not only the words "cherry" and "blossom" but also the phrase "cherry blossom," which is a type of flower. Finally, to test the model's performance without attention, we fine-tuned the GPT-4 model using the same procedure to obtain the GPT-4-fine-tuned model but added the prompt "loss of attention ability," resulting in the GPT-4-no-attention model. This model demonstrated the same logical ability as the GPT-4-fine-tuned model but incorrectly answered several questions that required long-term inference. This indicates that the attention cognitive ability of LLMs is essential for solving problems that require long-term reference, while strong logical ability is sufficient for solving short-term logical problems. Based on the performance results, the GPT-4 model also recognized the lack of attention ability in the GPT-4-no-attention model.

The result of the four tested models performance is shown in Table 1.

GPT-4 with attention demonstrates strong performance across the majority of the tasks, achieving the highest scores on 7 out of the 15 tasks. It particularly excels in tasks that require reasoning and understanding of complex concepts, such as Logical Deduction (5), Multistep Arithmetic, Navigation, Object Tracking, Logical Deduction (7), Penguins in a Table, and Reasoning about Colors. This suggests that the attention mechanism in GPT-4 with attention enables it to effectively handle tasks that require multi-step reasoning, numerical reasoning, and spatial reasoning. However, its performance on language-related tasks like Disambiguation and Hyperbaton is slightly lower compared to Vanilla GPT-4 and Claude-3.

Vanilla GPT-4 and Claude-3 show comparable performance across the tasks, with each model outperforming the other on specific tasks. Vanilla GPT-4 achieves the highest scores on Date Understanding, Hyperbaton, and Movie Recommendation, while Claude-3 performs best on Causal Judgement and Disambiguation. Both models demonstrate strong language understanding capabilities, as evident from their high scores on tasks like Salient Translation and Temporal Sequences. However, they lag behind GPT-4 with attention on tasks that require more complex reasoning, such as Logical Deduction and Multistep Arithmetic.

GPT-4 without attention consistently underperforms compared to the other three models across all tasks. Its scores are significantly lower than the other models on most tasks, with the exception of Salient Translation, where it achieves a score close to GPT-4 with attention. This suggests that the attention mechanism plays a crucial role in GPT-4's performance, and removing it severely impacts the model's ability to handle complex reasoning tasks. The low scores of GPT-4 without attention highlight the importance of the attention mechanism in enabling language models to effectively tackle challenging tasks that require reasoning and understanding of complex concepts.

In summary, the analysis of the model performance on the BIG-Bench Hard dataset reveals that GPT-4 with attention outperforms the other models on tasks that require complex reasoning and understanding, while Vanilla GPT-4 and Claude-3 demonstrate strong language understanding capabilities. GPT-4 without attention consistently underperforms, highlighting the importance of the attention mechanism in enabling language models to handle challenging tasks effectively. These results provide insights into the strengths and weaknesses of each model and underscore the significance of evaluating language models on a diverse set of complex tasks to comprehensively assess their capabilities.

5 Acknowledgments

We gratefully acknowledge the efforts of the students who participated in this exercise, and Dr. Yunhao Liu for hosting the project demonstrations and support.

Table 2: Test accuracy of each LLM

Model	Accuracy
GPT-3	75.0%
Claude-3	56.2%
GPT-4-with-Attention	81.2%
GPT-4-without-Attention	68.7%

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Alan Baddeley. 2003. Working memory: looking back and looking forward. *Nature reviews neuroscience* 4, 10 (2003), 829–839.

[3] Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences* 120, 6 (2023), e2218523120.

[4] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[5] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[7] Kahneman Daniel. 2017. *Thinking, fast and slow.*

[8] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* 27, 7 (2023), 597–600.

[9] Randall W Engle. 2002. Working memory capacity as executive attention. *Current directions in psychological science* 11, 1 (2002), 19–23.

[10] Martha J Farah. 2000. *The cognitive neuroscience of vision.* Blackwell Publishing.

[11] Lesley K Fellows. 2004. The cognitive neuroscience of human decision making: a review and conceptual framework. *Behavioral and cognitive neuroscience reviews* 3, 3 (2004), 159–172.

[12] John DE Gabrieli. 1998. Cognitive neuroscience of human memory. *Annual review of psychology* 49, 1 (1998), 87–115.

[13] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3, 10 (2023), 833–838.

[14] Jason Holmes, Zhengliang Liu, Lian Zhang, Yuzhen Ding, Terence T Sio, Lisa A McGee, Jonathan B Ashman, Xiang Li, Tianming Liu, Jiajian Shen, et al. 2023. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Frontiers in Oncology* 13 (2023), 1219326.

[15] Keith J Holyoak and Robert G Morrison. 2005. *The Cambridge handbook of thinking and reasoning.* Cambridge University Press.

[16] Philip Johnson-Laird. 2008. *How we reason.* Oxford University Press.

[17] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, Vol. 1. 2.

[18] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).

[19] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[20] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. GPT understands, too. *AI Open* (2023).

[21] Michael I Posner, Steven E Petersen, et al. 1990. The attention system of the human brain. *Annual review of neuroscience* 13, 1 (1990), 25–42.

[22] Yanina Prystauka, Vincent DeLuca, Alicia Luque, Toms Voits, and Jason Rothman. 2023. Cognitive Neuroscience Perspectives on Language Acquisition and Processing. , 1613 pages.

[23] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[24] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[25] Denis Rothman. 2022. *Transformers for Natural Language Processing: Build, train, and fine-tune deep neural network architectures for NLP with Python, Hugging Face, and OpenAI's GPT-3, ChatGPT, and GPT-4.* Packt Publishing Ltd.

[26] Edward E Smith and Stephen Michael Kosslyn. 2007. Cognitive psychology: Mind and brain. *(No Title)* (2007).

[27] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).

[28] Gaurav Suri, Lily R Slater, Ali Ziaee, and Morgan Nguyen. 2024. Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General* (2024).

[29] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261* (2022).

[30] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine* 29, 8 (2023), 1930–1940.

[31] A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* (2017).

[32] Chaozheng Wang, Yuanhang Yang, Cuiyun Gao, Yun Peng, Hongyu Zhang, and Michael R Lyu. 2022. No more fine-tuning? an experimental evaluation of prompt tuning in code intelligence. In *Proceedings of the 30th ACM joint European software engineering conference and symposium on the foundations of software engineering*. 382–394.

[33] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564* (2023).

A Online Resources

https://github.com/xiaoyanLi629/RMELS2024

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009